

A comparison of multiple imputation methods for bivariate hierarchical outcomes

K. DiazOrdaz, M. G. Kenward, M. Gomes, R. Grieve

July 18, 2014

Abstract

Missing observations are common in cluster randomised trials. Approaches taken to handling such missing data include: complete case analysis, single-level multiple imputation that ignores the clustering, multiple imputation with a fixed effect for each cluster and multilevel multiple imputation.

We conducted a simulation study to assess the performance of these approaches, in terms of confidence interval coverage and empirical bias in the estimated treatment effects. Missing-at-random clustered data scenarios were simulated following a full-factorial design. An Analysis of Variance was carried out to study the influence of the simulation factors on each performance measure.

When the randomised treatment arm was associated with missingness, complete case analysis resulted in biased treatment effect estimates. Across all the missing data mechanisms considered, the multiple imputation methods provided estimators with negligible bias. Confidence interval coverage was generally in excess of nominal levels (up to 99.8%) following fixed-effects multiple imputation, and too low following single-level multiple imputation. Multilevel multiple imputation led to coverage levels of approximately 95% throughout.

The approach to handling missing data was the most influential factor on the bias and coverage. Within each method, the most important factors were the number and size of clusters, and the intraclass correlation coefficient.

1 Introduction

In cluster randomised trials, the unit of random allocation is a group of individuals (e.g. a school or a hospital) rather than the individual subjects. It is a common study design in the health and social sciences, especially for evaluations of interventions that operate at a group level, manipulate the socio-physical environment, or cannot be delivered at an individual level. It is well-known that observations within each cluster are correlated (Cornfield, 1978) and that analyses that ignore this homogeneity within clusters can result in overestimation of the precision of the treatment effects, possibly leading to inappropriate inferences being drawn. Appropriate statistical techniques for cluster randomised trials are well developed and include mixed models and generalised estimating equations (Donner and Klar, 2000).

A common problem that compromises the validity of the results is that of missing data. The validity of inferences from incomplete data depends on the process that leads to data being missing, the so-called missing data mechanism, also known as *missingness mechanism* or *missing data process* (Molenberghs and Kenward, 2007, Section 3.2). The missing data mechanism is characterised by the conditional distribution of the probability of missingness, given the data. Rubin (1987) proposed a classification of the missing data mechanisms, according to the assumed model for the probability of non-response. A process is said to be Missing Completely at Random (MCAR) if the probability of non-response is completely independent of the measurement process. A process is classified as Missing at Random (MAR) if the probability of non-response is conditionally independent of the unobserved data given the observed data. Processes that are neither MCAR nor MAR are called missing not at random (MNAR).

For missing data mechanisms that satisfy MAR, valid inferences can be obtained using likelihood-based or Bayesian analyses of the complete cases (Molenberghs and Kenward, 2007, Part III). However,

moment-based estimators, such as those that use generalised estimating equations are, without special modification, only valid with more stringent conditions about the missing data mechanism, namely that the data are MCAR.

A commonly used approach to obtain valid inferences for incomplete data under the MAR assumption is Multiple Imputation (MI) (Rubin, 1987). In some circumstances, essentially when the analysis and imputation models coincide, MI principally replicates a likelihood analysis. However, an advantage of MI is that unlike conventional likelihood analyses, it can incorporate so-called auxiliary variables that are not included in the analysis model, but which are related to both the missing values and to the probability of observations being missing. Incorporating such auxiliary variables makes the underlying MAR assumption more plausible.

From a theoretical perspective, it is known that for cluster randomised trials, the imputation method should accommodate the multilevel structure of the data. A failure to do this may lead to invalid inferences (Schafer and Yucel, 2002). However, multilevel multiple imputation is not yet available as a standard implementation in commonly used statistical packages, although particular routines are available, for example, Schafer (2001); Carpenter *et al.* (2011). Hence, analyses using MI in the cluster randomised trials settings commonly avoid such imputation strategies, and use instead imputation methods that ignore the clustering (Díaz-Ordaz *et al.*, 2014). An alternative approach that has been previously recommended in the literature is including the cluster as a fixed effect (White *et al.*, 2011; Graham, 2009). This has the advantage of being easily implemented in widely available MI software.

Previous simulation-based comparisons of the alternative methods have been presented by Taljaard *et al.* (2008) and Andridge (2011), using a single missing outcome and missing data mechanisms under both MCAR and MAR dependent on individual-level variables. In the present study, we consider situations where we wish to simultaneously analyse several responses as functions of the explanatory variables using random effects models. Examples of such models include cost-effectiveness analysis, where policy-makers require an estimate of the alternative treatments on the joint distribution of costs and health outcomes. We focus on bivariate responses with missing data, but the conclusions for univariate or multivariate outcomes follow directly from these. The simulation does not consider missing covariate data nor data which are missing not at random.

The aim of this paper is two-fold. The first is to investigate the relative performance of different multiple imputation strategies for handling missing bivariate outcome data in cluster randomised trials, over a wide range of missingness mechanisms that are dependent on individual and cluster-level variables. The second is to explore the effect of different trial characteristics, such as number and size of the clusters and level of clustering, on the performance of the MI estimator in finite samples given one of the above MI strategies. A simulation study with a factorial design is used for this.

The remainder of this paper is organised as follows. In the next section, we provide some details on the alternative MI methods. In Section 3, we describe the simulation study and its analysis, in particular making use of the factorial structure through analysis of variance type procedures. In Section 4, we report a selection of results from simulated scenarios. We close with a few points of interpretation and discussion in Section 5.

2 Multiple Imputation

Multiple imputation breaks down the analysis of incomplete data into a number of steps. We first need to distinguish between two statistical models. The first is the analysis model that would have been used had the data been complete. This is called the *substantive model* or *model of interest*. The second model, called the *imputation model*, is used to describe the conditional distribution of the missing data given the observed. For hierarchical data, this conditional distribution must reflect the multilevel nature of the data.

The MI algorithm proceeds by fitting the imputation model to the observed data and taking Bayesian draws from the posterior distribution of its model parameters. Missing data are then imputed from the imputation model, using the parameters previously drawn. These steps are repeated a fixed M number

of times, to obtain M completed data sets. The substantive model is then fitted to the multiple data sets separately, producing M sets of parameter and covariance estimates which are combined using Rubin's formulae (Rubin, 1987) to produce a single MI estimate of the substantive model parameters and associated covariance matrix.

Under the MAR assumption, this will produce consistent estimators and, in the absence of auxiliary variables, is asymptotically (as M increases) equivalent to maximum likelihood (Little and Rubin, 2002; Schafer, 1997).

Sampling from the approximate predictive distribution of the missing data as described above can be performed in several ways. Two broad approaches can be identified; the first approach jointly models incomplete variables, by sampling from an underlying joint predictive distribution (Schafer, 1997; Goldstein *et al.*, 2009). In the second approach, referred to as full-conditional specification (FCS) or *chained equations*, draws from the joint distribution are approximated using a sampler consisting of a set of univariate models for each incomplete variable conditional on all the other variables (van Buuren, 2012).

In the simulations presented here, both approaches are used. For single-level imputation and fixed cluster effects models, which are also essentially single-level, the FCS method is used, as implemented in the MICE package in R (van Buuren and Groothuis-Oudshoorn, 2011). The FCS approach is not well-suited to proper multilevel MI and so, for these imputations, Schafer's PAN package is used (Schafer, 2001).

Having outlined the generic MI procedure, we now set out the details of the relevant imputation models to be compared here.

Let $Y_{1,ij}$ and $Y_{2,ij}$ be the two continuous outcomes with missing data, corresponding to the i -th individual in cluster j of a two-arm cluster trial. Let treatment allocation be represented by $k = 1$, if the cluster is allocated to intervention, and 0 otherwise. Let \mathbf{X}_{ijk} denote the matrix of all auxiliary variables (assumed to be fully observed), including individual and cluster-level variables.

The imputation models compared here express $(Y_{1,ijk}, Y_{2,ijk})$ as a function of the grand mean in treatment arm k ($\nu_{\ell,0k}$, for $\ell = \{1, 2\}$), the auxiliary variables, and error terms $(e_{1,ijk}, e_{2,ijk})$.

The single-level imputation model (SMI) can be written as:

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X}_{ijk}\nu_{1,X} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X}_{ijk}\nu_{2,X} + e_{2,ijk} \end{aligned} \quad \begin{pmatrix} e_{1,ijk} \\ e_{2,ijk} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Omega}_1)$$

where $\nu_{\ell,X}$ is the vector of regression coefficients, and $\mathbf{\Omega}_1$ is the individual-level variance-covariance matrix. With single-level MI, the imputed values are drawn from the conditional distribution of the missing observations given the observed data, ignoring any dependency between observations within a cluster not explained by the cluster-level auxiliary variables included in the model. Therefore, the single-level imputation model does not properly represent the conditional distribution of the missing data given the observed data.

Two imputation models have been used to incorporate the effect of clustering. Firstly, we include a cluster fixed-effect in the imputation model (denoted FMI):

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X}_{ijk}\nu_{1,X} + \sum_{j=1}^{J-1} \beta_{1,j}I_{ij} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X}_{ijk}\nu_{2,X} + \sum_{j=1}^{J-1} \beta_{2,j}I_{ij} + e_{2,ijk} \end{aligned}$$

where I_{ij} is the indicator variable for cluster j , so that $I_{ij} = 1$ if the observation i belongs to cluster j and the error term $(e_{1,ijk}, e_{2,ijk})$ is assumed to be bivariate normal as before. This model allows a different intercept for each cluster within treatment group k . Missing outcomes will be imputed from the conditional normal distribution given the other outcome, if observed, and the auxiliary variables, which must all be at

the individual level, with a mean determined by the fixed-effect for that cluster.

Secondly, we include a *random effects* for clustering in the imputation model (denoted MMI):

$$\begin{aligned} Y_{1,ijk} &= \nu_{1,0k} + \mathbf{X}_{ijk}\nu_{1,X} + b_{1,j} + e_{1,ijk} \\ Y_{2,ijk} &= \nu_{2,0k} + \mathbf{X}_{ijk}\nu_{2,X} + b_{2,j} + e_{2,ijk} \end{aligned} \quad \begin{pmatrix} b_{1,j} \\ b_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Omega}_2)$$

where $\mathbf{\Omega}_2$ is the cluster-level variance-covariance matrix and the individual-level residuals $(e_{1,ijk}, e_{2,ijk})$ are assumed normally distributed independently of $(b_{1,j}, b_{2,j})$, the cluster random-effects.

Finally, complete-case analysis (CCA) is also included in our simulations, for comparative purposes.

3 Simulation study

A simulation study following a full factorial design was conducted comparing the performance of the methods considered here. The simulation steps proceeded as follows: data generation, application of a missing data mechanism, and estimation and inference for the treatment effect from the analysis after handling (or ignoring) the missing data. Finally, the behaviour of the treatment effect estimator is examined according to our chosen performance measures.

3.1 Data generation

For each subject i in cluster j , standard normal individual-level covariate X_i and cluster-level variable W_j were generated. Bivariate normal outcome data $(Y_{1,ijk}, Y_{2,ijk})$ were then generated depending on these covariates, separately in each treatment arm $k = 0, 1$.

The level of clustering, quantified by the intraclass correlation coefficient (ICC), was allowed to vary according to the levels set out in Table 1. The number and size of clusters were also varied, while maintaining the same overall sample size ($S = 500$). Three different types of cluster randomised trial design were considered: (i) large number of clusters ($J = 50$) and few individuals per cluster ($n_j = 10$); (ii) small number of clusters ($J = 10$) and large cluster size ($n_j = 50$); (iii) moderate number of clusters (30) and variable number of individuals per cluster. For these scenarios, cluster size n was assumed to follow a Gamma distribution, with mean 20 and coefficient of variation $cv = \frac{SD(n)}{E(n)} = 0.5$.

3.2 Missing data mechanisms

To generate the missing data under the Missing-at-Random assumption, we used four different missing data mechanisms, where the probability of non-response, denoted by $\pi_{\ell,ijk}$, was such that the non-response indicator $R_{\ell,ijk} \sim \text{Bern}(\pi_{\ell,ijk})$, depends on X_i or/and W_j , as displayed also in Table 1. The coefficient η represents the strength of association between the covariates and non-response indicator $R_{\ell,ijk}$. We adjusted α_0 empirically to achieve the required expected probability of missing.

We selected other factors which were anticipated to have an impact on the performance of the approaches for handling missing data, based on previous literature (Rubin, 1987; Taljaard *et al.*, 2008; Andridge, 2011; Carpenter and Kenward, 2013). For the first three missing data mechanisms, the same factors and levels were used for both randomised treatment arms. These are reported in Table 1.

We assumed that for both outcomes, individual and cluster level covariates have the same level of association η with the non-response indicator, and thus we drop the indexes ℓ and X, W . However, for the last of our missingness mechanisms, we allowed η to differ between treatment arms, with two settings, and these are presented in Table 2, together with the probabilities of non-response, which also differ across treatment arms.

Non-response rates were chosen to minimise the number of clusters with one or both outcomes completely missing, as whole cluster non-response raises other issues not dealt here.

For each simulated dataset, non-response indicators $R_{\ell,ijk}$ for each outcome were independently drawn from a Bernoulli distribution with probabilities $\pi_{\ell,ijk}$ as specified in Table 1. Missing values were then generated to create the *observed* data set.

3.3 Substantive model

We focus on likelihood-based methods for the substantive model, rather than estimating equations. Because the intervention effect lies at the cluster level, likelihood-based methods that acknowledge the clustering must use *random* cluster effects; fixed cluster effects would absorb all the information on the intervention effects.

Hence, the substantive model is a bivariate Gaussian random-effects model where the only explanatory variable is treatment. Let the cluster-level random effects be represented by the latent variables $u_{1,j}$ and $u_{2,j}$. The model can be written as follows

$$\begin{aligned} Y_{1,ij} &= \beta_{1,0} + \beta_1 t_j + u_{1,j} + e_{1,ij} \\ Y_{2,ij} &= \beta_{2,0} + \beta_2 t_j + u_{2,j} + e_{1,ij} \end{aligned} \quad (1)$$

where β_1 and β_2 represent the treatment effect on the corresponding outcome. The error term $(e_{1,ij}, e_{2,ij})$ and the cluster effects are assumed to be normally distributed:

$$\begin{pmatrix} e_{1,ij} \\ e_{2,ij} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right] \text{ and } \begin{pmatrix} u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \phi\tau_1\tau_2 \\ \phi\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right]$$

where σ_1, σ_2 are the individual-level standard errors, ρ is the individual-level correlation between Y_1 and Y_2 and τ_1, τ_2 , and ϕ are the standard errors and correlation of the two cluster random effects, respectively.

3.4 Implementation

The number of imputations M was set at 10. After imputation, for which the two covariates were used as auxiliary variables, the substantive model, equation (1), was applied to each multiply imputed dataset to estimate treatment effect on Y_1 and Y_2 simultaneously. The estimates obtained using the analysis model in each of the M multiply imputed sets were then combined using Rubin's rules.

For each scenario, the whole simulation procedure (data generation, imposing missing values, imputation, analysing each of the imputed datasets using the substantive model, and combining the resulting treatment effect estimates using Rubin's rules) was performed on $N = 1000$ datasets to capture the behavior in repeated samples.

3.5 Performance criteria

Let θ denote the true treatment effect parameter, and $\hat{\theta}_l$ the estimate obtained in the $l = 1, \dots, N$ replicated dataset. The following criteria were used to measure the performance of the different MI strategies.

1. Confidence interval coverage rate (CR): The percentage of times that the true parameter value is covered in the 95% confidence interval.
2. Empirical bias, $B = \frac{1}{N} \sum_{l=1}^N \hat{\theta}_l - \theta$
3. Root-mean-square error (RMSE) $\sqrt{\frac{1}{N} \sum_{l=1}^N (\hat{\theta}_l - \theta)^2}$
4. Average width of confidence interval (AW): The distance between the average lower and upper confidence interval limits across N confidence intervals.

The performance of a procedure is regarded as poor if its coverage drops below 90% (Collins *et al.*, 2013). If the procedure results in CRs that are close to 100% extra caution should be taken when using that procedure (Yucel *et al.*, 2010). A CR close to the nominal value, along with narrow confidence intervals translates into greater accuracy and higher power.

3.6 Analysis of the simulation results

The factorial structure of the simulation design was exploited through the use of analysis of variance (ANOVA) summaries to isolate key factors that are associated with large impact on the performance of the multiple imputation procedures.

ANOVA was carried out on each performance measure for each outcome including all main effects and interactions amongst factors up to 4-way interactions (the rest constituting the “residual” degrees of freedom). The relative size of the F-statistics derived from the ANOVA was used as an indicator of the influence of that factor (or interaction of factors) on the particular performance measure. The F-statistic can be thought of as the ratio of variability explained/variability unexplained by the model. Multivariate ANOVA (MANOVA) was also used to study the impact of each factor on the overall performance. For this, the Wilks Lambda statistic was calculated to obtain an approximate F-statistic.

These F-statistics are not used in an inferential way, and no distributional assumptions are being made; instead they are used as a descriptive measure of influence. To help visualise this, we normalised the value of these F-statistics by dividing each F-value by the largest of those obtained in for each performance measure. For bias and confidence interval coverage, we calculated the proportion of simulated scenarios where the performance measure in turn was deemed unsatisfactory, that is bias which is larger than 1.96 times the Monte Carlo error, and confidence interval coverage which is either lower than 90% or higher than 97%. We then multiply the re-normalised F-values by this proportion and plotted the resulting number by performance measure and missing data approach.

Since one of our aims is to establish what influences performance within each MI method, these analyses were also performed stratified by MI method. In addition, we report the range of the percentage bias and coverage rate over each of the simulation factors and plot the distribution of these performance measures stratified by MI method and missing data mechanism.

4 Results

Without loss of generality, we present here the results corresponding to bias and coverage for treatment effect estimates on Y_1 . The corresponding results for Y_2 are available from the corresponding author upon request.

The distribution of bias and coverage rate by MI method and missing mechanism are shown in Figure 1. We observe that, for the first three missing data mechanisms studied (see Table 1), where the missing data mechanism was not dependent on treatment arm, all approaches resulted in unbiased estimates across most of the scenarios. This is in line with theoretical results, as the variables associated with missingness are not associated with the treatment effect. However, for the scenario when the missing mechanism is differential by treatment arm, the CCA produced substantially biased estimates across the scenarios considered. The corresponding results for the MI estimates show none to negligible bias: in general less than 3.5%, and mostly within Monte Carlo error limits. This is reported in Table 5.

However, the alternative MI strategies resulted in very different variance estimates, and coverage rates. Table 5 reports the coverage rate and average width again for the scenarios where the missingness mechanism was different by treatment arms. Similar tables corresponding to the other missingness mechanism are reported in the Supplementary file. In particular, the single-level MI resulted in substantial under-coverage for scenarios with high ICCs (0.20 and above). The number and size of clusters also appear to be factors associated with low coverage rate. For scenarios where the number of clusters is relatively low ($J = 5$

per arm), multilevel and single-level MI both have low coverage for the estimated treatment effect. Fixed-effects MI results in over-conservative coverage for a range of scenarios, especially those corresponding to missing data mechanisms that depend only on a cluster-level variable (Table 5 in the Supplementary File) and where the ICCs are moderate to small. Across all methods and scenarios considered, the accuracy is similar, i.e. comparable RMSE, however FMI is systematically inefficient, wider confidence intervals are obtained using FMI compared to those obtained using either SMI or MMI, even when estimates are unbiased and coverage is acceptable.

The ANOVA results confirm that the most influential factor on the validity of inferences drawn from missing data is the method chosen to handle these. Within each MI method, the ANOVA results on bias and coverage rate are reported graphically in Figure 2. Terms with very small normalised F-statistic value are not plotted. For CCA, the most influential factor for the substantial empirical bias is the missing data mechanism. For SMI and MMI empirical bias is low, and the strength of association between the covariates and the non-response indicator is almost as influential as the missing data mechanism. For FMI, while bias is again low, the ANOVA suggests that ICC and the number and size of the clusters are the determining factors affecting bias.

The corresponding figure for coverage rate shows that most influential factors are the level of clustering, measured by the ICC and the number and size of the clusters (denoted in the figures as *Design*). Nevertheless, for CCA, the most influential factor is the missing data mechanism. As the relative height of the bars show, the method which most consistently achieves coverage rates close to the nominal is MMI, as it has the smallest proportion of scenarios with over or under-coverage, with only 8 scenarios out of the total 192 resulting in CR lower than 90% and higher than 97%.

5 Discussion

In this simulation study, we compared the performance of single, multilevel and fixed-effects MI for handling missing data in cluster randomised trials. The full-factorial nature of our simulation study enabled us to establish which characteristics have the greatest influence on the performance of the alternative methods for handling missing data considered here.

In our simulations, which assumed the data were MAR throughout, bias was a serious problem for the complete case analysis when the missingness mechanism was differential by treatment arm, while all MI methods resulted in unbiased treatment estimates. The main difference amongst the three MI procedures is in how variability is incorporated into the imputations. Single-level MI resulted in low ($< 90\%$) coverage rate across most scenarios, in particular when the ICCs exceeded 0.05 and there were few clusters. Fixed-effects MI produced overly conservative coverage ($> 97\%$), especially when there were small ICCs and more than 30 clusters. This finding reflects the way these two approaches accommodate the between-cluster variance. Under single-level MI, the between-cluster variance is set to zero, whereas with the fixed-effect MI this variance is unbounded in the sense that the behaviour of one estimated cluster effect is unrelated, or unconstrained, by the behaviour of any of the others. Indeed, including cluster as a fixed-effect in the imputation model represents the limiting case where the proportion of variability at the cluster-level tends to one and does not properly capture the conditional distribution of the missing data given the observed. It cannot be used when cluster-level variables need to be imputed, and appears to perform worse when the missing data mechanism is driven by a cluster-level covariate, which cannot be explicitly included in the imputation model.

By contrast, multilevel MI models the correlation in the data appropriately, producing coverage rates close to the nominal level. This consistent performance across the varying sample sizes is indicative of acceptable finite sample properties. Moreover, multilevel MI is compatible with the substantive model, which uses cluster random effects, and the imputation model can include auxiliary variables at both the individual and the cluster-level, thus increasing the plausibility of the MAR assumptions.

Our findings underscore the importance of selecting an imputation model with a compatible, multilevel structure to that of the substantive models, and corresponds to those from previous studies which have

compared single and multilevel MI in settings with hierarchical data. For example, Taljaard *et al.* (2008) found that single-level MI results in excessive Type I errors in settings where data were MCAR. Our study also extends the results of Andridge (2011), who found that including cluster as a fixed effect in the imputation model overestimates the variance, especially when ICCs are low, and there are few clusters.

The results presented in this study could potentially be extended to other situations. Our imputation and substantive models match exactly the data generating process, but previous simulation studies (Schafer, 1997; Yucel *et al.*, 2010) have shown that MI is fairly robust to distributional misspecification of the imputation model. Also, for simplicity, we assumed the missing data mechanism is MAR throughout. An interesting extension would be to explore MNAR mechanisms, especially those when the cluster random effect is driving the missingness. Other potential extensions relate to situations where there is cluster non-response. In both situations, multilevel MI could provide a flexible route for investigating sensitivity to alternative MNAR mechanisms and cluster drop-out (Carpenter and Kenward, 2013, Chapter 10).

Acknowledgments

The authors will like to thank James Carpenter for helpful discussions. KDO is funded by an MRC Career development award in Biostatistics, MG by an MRC Early Career fellowship in Economics of Health, and RG by a senior research fellowship from NIHR.

Tables and Figures

Table 1: Factors and their chosen levels that differ across scenarios for missingness mechanisms which do not differ by treatment arm

Factor	Levels	Values
ICC ₁ and ICC ₂	low	(0.01, 0.01)
	moderate	(0.20, 0.05)
	high	(0.20, 0.20)
	differential by outcome	(0.60, 0.01)
Cluster design	many small clusters	$J = 50, n_j = 10$
	few large clusters	$J = 10, n_j = 50$
	unbalanced	$J = 30$, variable size
Missingness mechanism	Individual covariate	$\text{logit } \pi_{\ell,ij} = \alpha_0 + \eta X_i$
	Cluster covariate	$\text{logit } \pi_{\ell,ij} = \alpha_0 + \eta W_j$
	Both	$\text{logit } \pi_{\ell,ij} = \alpha_0 + \eta X_i + \eta W_j$
	Differential by treatment	$\text{logit } \pi_{\ell,ijk} = \alpha_{0k} + \eta_k X_{ij} + \eta_k W_j$
Association between covariates and missingness	low	$\eta = 1$
	high	$\eta = 2$
Probability of Non-response	equal	20%
	different by outcome	30% for $Y_{1,ij}$; 10% for $Y_{2,ij}$

Table 2: Parameter values for settings where the missingness mechanism is differential by treatment arm.

Level of association	Arm	Association with missingness	Probability of non-response		
			equal	Differential by outcome	
low	Control	$\eta_0 = 1$	20%	30%	10%
	Intervention	$\eta_1 = 2$	35%	45%	20%
high	Control	$\eta_0 = 1.5$	10%	15%	10%
	Intervention	$\eta_1 = 3$	30 %	35%	30%

The numbers in italics are not simulation parameters, but the approximate empirical rates of non-response obtained after setting α_0 .

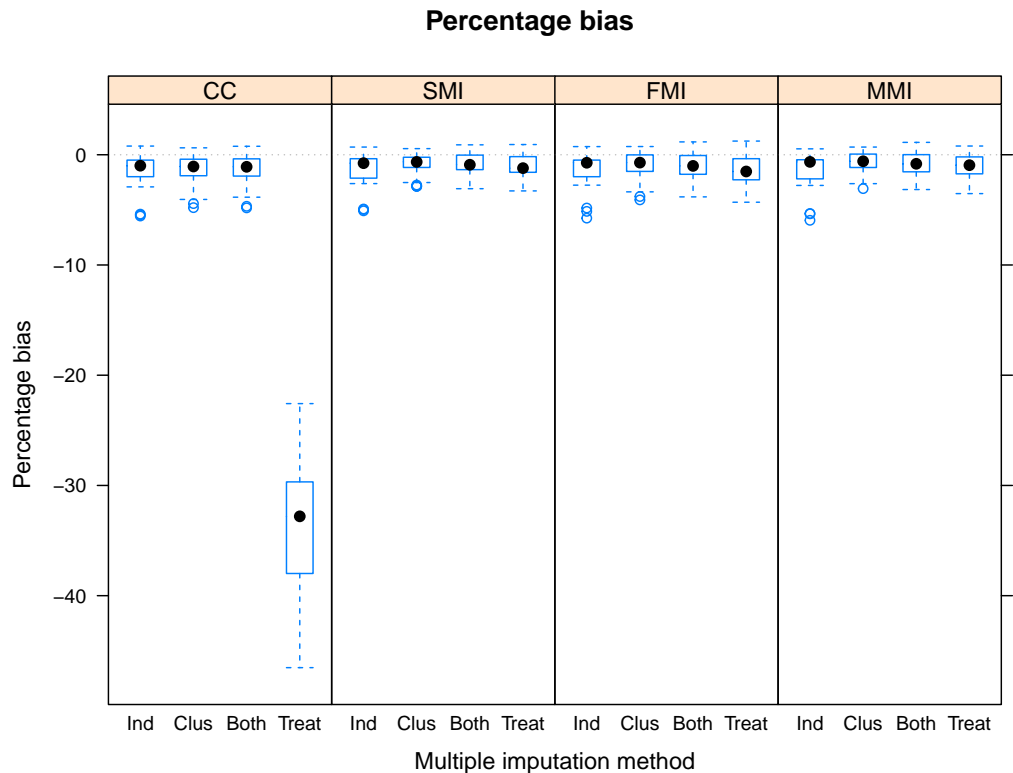
Table 3: Percentage bias for the estimated treatment effect on Y_1 for scenarios corresponding to missingness mechanism is differential by treatment

Design	η	Missingness	ICC	CCA	SMI	FMI	MMI
$J = 50, n_j = 10$	Low	.20,.20	0.01, 0.01	-24.8	-1.4	-0.8	-0.8
			0.20, 0.05	-32.9	-1.5	-1.3	-1.0
			0.20, 0.20	-33.1	-1.6	-1.3	-1.0
			0.60, 0.01	-38.7	-1.4	-2.1	-1.4
		.30,.10	0.01, 0.01	-23.2	-1.3	-0.7	-0.2
			0.20, 0.05	-31.0	-1.6	-1.7	-0.3
			0.20, 0.20	-31.1	-1.6	-1.7	-0.4
			0.60, 0.01	-35.9	-1.8	-3.5	-0.5
		High	.20,.20	0.01, 0.01	-28.2	-1.7	-2.3
				0.20, 0.05	-37.5	-1.7	-3.0
				0.20, 0.20	-37.9	-1.9	-3.0
				0.60, 0.01	-43.4	-1.5	-4.2
			.30,.10	0.01, 0.01	-29.2	-1.3	-1.1
				0.20, 0.05	-39.3	-1.5	-2.3
				0.20, 0.20	-39.7	-1.6	-2.3
				0.60, 0.01	-46.5	-1.4	-4.3
	High	.20,.20	0.01, 0.01	-25.2	0.1	-0.2	-0.9
			0.20, 0.05	-31.1	-1.0	-1.5	-1.8
			0.20, 0.20	-31.5	-1.0	-1.5	-1.7
			0.60, 0.01	-32.7	-2.8	-3.7	-3.5
		.30,.10	0.01, 0.01	-24.5	-0.1	-0.4	-0.9
			0.20, 0.05	-30.2	-1.3	-1.7	-1.9
			0.20, 0.20	-30.7	-1.3	-1.7	-1.9
			0.60, 0.01	-31.7	-3.3	-4.0	-3.5
		High	.20,.20	0.01, 0.01	-29.2	0.0	-0.4
				0.20, 0.05	-36.5	-1.1	-1.7
				0.20, 0.20	-37.0	-1.2	-1.7
				0.60, 0.01	-38.5	-2.9	-4.0
			.30,.10	0.01, 0.01	-31.1	0.3	-1.5
				0.20, 0.05	-38.9	-0.7	-1.7
				0.20, 0.20	-39.5	-0.8	-1.6
				0.60, 0.01	-41.0	-2.3	-4.3
$J = 10, n_j = 50$	Low	.20,.20	0.01, 0.01	-23.1	0.9	1.2	0.4
			0.20, 0.05	-30.3	0.4	0.6	0.0
			0.20, 0.20	-30.6	0.3	0.5	-0.1
			0.60, 0.01	-33.7	-0.6	-0.8	-1.1
		.30,.10	0.01, 0.01	-22.6	0.5	1.2	0.4
			0.20, 0.05	-29.6	-0.4	0.2	-0.2
			0.20, 0.20	-29.8	-0.5	0.3	-0.3
			0.60, 0.01	-33.0	-2.0	-1.5	-1.2
		High	.20,.20	0.01, 0.01	-26.8	0.8	0.4
				0.20, 0.05	-35.4	0.3	-0.3
				0.20, 0.20	-35.8	0.2	-0.3
				0.60, 0.01	-39.6	-0.5	-1.8
			.30,.10	0.01, 0.01	-28.3	0.7	0.5
				0.20, 0.05	-37.6	-0.2	-0.6
				0.20, 0.20	-38.1	-0.4	-0.7
				0.60, 0.01	-42.6	-1.8	-2.7
$J = 30, \text{unbalanced}$	Low	.20,.20	0.01, 0.01	-23.1	0.9	1.2	0.4
			0.20, 0.05	-30.3	0.4	0.6	0.0
			0.20, 0.20	-30.6	0.3	0.5	-0.1
			0.60, 0.01	-33.7	-0.6	-0.8	-1.1
		.30,.10	0.01, 0.01	-22.6	0.5	1.2	0.4
			0.20, 0.05	-29.6	-0.4	0.2	-0.2
			0.20, 0.20	-29.8	-0.5	0.3	-0.3
			0.60, 0.01	-33.0	-2.0	-1.5	-1.2
		High	.20,.20	0.01, 0.01	-26.8	0.8	0.4
				0.20, 0.05	-35.4	0.3	-0.3
				0.20, 0.20	-35.8	0.2	-0.3
				0.60, 0.01	-39.6	-0.5	-1.8
			.30,.10	0.01, 0.01	-28.3	0.7	0.5
				0.20, 0.05	-37.6	-0.2	-0.6
				0.20, 0.20	-38.1	-0.4	-0.7
				0.60, 0.01	-42.6	-1.8	-2.7

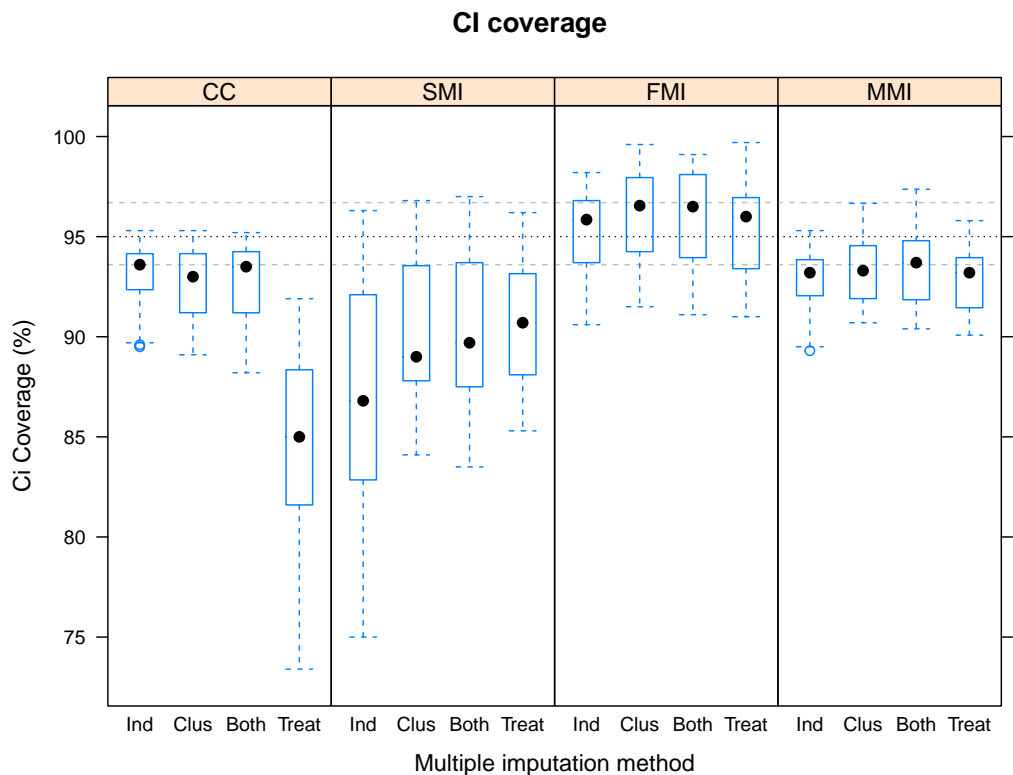
Table 4: Coverage rate (CR) and average width (AW) corresponding to confidence interval of the treatment effect estimate, when missingness is differential by treatment arm.

Design	η	Missingness	ICC	CCA		SMI		FMI		MMI	
				CR	AW	CR	AW	CR	AW	CR	AW
$J = 50,$ $n_j = 10$	Low	.20,.20	0.01, 0.01	81.2	18.6	95.5	17.9	98.8	22.3	94.9	17.5
			0.20, 0.05	84.7	27.6	92.6	26.2	96.5	31.2	93.1	27.4
			0.20, 0.20	83.9	27.7	92.8	26.3	96.3	31.2	93.1	27.2
			0.60, 0.01	91.3	56.0	90.7	51.4	95.1	58.7	94.4	56.9
		.30,.10	0.01, 0.01	82.2	18.7	95.5	19.4	99.7	26.1	94.9	18.9
			0.20, 0.05	85.2	27.7	92.5	26.9	97.5	34.0	93.0	28.2
			0.20, 0.20	84.6	27.7	92.6	27.0	97.5	34.0	92.4	27.9
			0.60, 0.01	91.7	56.0	90.7	50.8	95.2	60.1	93.5	57.5
	High	.20,.20	0.01, 0.01	75.1	17.4	95.6	17.3	98.7	21.4	95.6	17.2
			0.20, 0.05	81.2	26.7	91.8	26.2	96.8	30.6	93.8	27.4
			0.20, 0.20	81.2	26.6	91.9	26.2	96.8	30.6	93.7	27.2
			0.60, 0.01	89.7	55.2	91.0	52.3	94.6	58.4	94.4	56.9
		.30,.10	0.01, 0.01	73.7	17.8	96.2	18.5	99.1	23.4	95.8	18.1
			0.20, 0.05	78.4	26.8	92.2	26.8	96.8	32.0	93.2	28.0
			0.20, 0.20	78.6	26.8	92.3	26.8	96.8	32.1	93.6	27.8
			0.60, 0.01	89.3	55.3	90.4	52.2	95.0	59.1	94.4	57.4
$J = 10,$ $n_j = 50$	Low	.20,.20	0.01, 0.01	82.0	20.3	94.8	20.2	96.5	22.6	95.6	20.2
			0.20, 0.05	87.4	48.1	87.4	47.0	92.4	53.1	91.0	51.1
			0.20, 0.20	87.2	48.4	87.8	47.0	92.6	53.1	90.9	51.1
			0.60, 0.01	89.9	116.1	87.5	107.3	91.1	121.5	90.9	120.2
		.30,.10	0.01, 0.01	83.2	20.3	94.8	21.8	97.2	25.4	95.3	21.7
			0.20, 0.05	86.6	47.9	87.2	46.3	92.1	54.3	90.3	51.2
			0.20, 0.20	87.7	48.2	87.2	46.4	92.4	54.1	90.2	51.0
			0.60, 0.01	89.0	115.7	85.3	103.7	91.0	121.5	90.7	120.0
	High	.20,.20	0.01, 0.01	75.6	19.4	94.8	19.8	96.1	21.9	94.1	20.0
			0.20, 0.05	85.4	47.3	88.9	47.9	92.0	52.7	91.4	51.3
			0.20, 0.20	85.3	47.5	88.9	47.9	92.1	52.6	91.4	51.3
			0.60, 0.01	89.7	115.0	87.9	110.2	91.2	121.0	91.0	120.2
		.30,.10	0.01, 0.01	74.3	20.0	94.6	20.9	97.1	27.0	95.1	21.0
			0.20, 0.05	85.1	47.6	87.6	47.7	92.8	54.3	90.1	51.7
			0.20, 0.20	84.9	47.8	87.5	47.7	92.4	53.8	90.2	51.6
			0.60, 0.01	89.1	115.0	86.9	108.4	91.0	123.7	90.8	120.9
$J = 30,$ unbalanced	Low	.20,.20	0.01, 0.01	81.0	17.7	93.9	17.0	97.4	21.0	93.2	16.9
			0.20, 0.05	85.9	32.1	89.9	30.3	95.7	36.1	93.0	32.9
			0.20, 0.20	85.9	32.1	89.7	30.3	96.0	35.9	93.1	32.7
			0.60, 0.01	91.2	70.3	89.6	63.7	94.3	74.0	93.5	72.4
		.30,.10	0.01, 0.01	82.6	17.8	93.8	18.4	98.5	24.6	93.9	18.3
			0.20, 0.05	85.7	32.1	88.0	30.4	96.4	38.3	92.2	33.4
			0.20, 0.20	85.5	32.1	88.2	30.4	96.2	38.1	91.5	33.1
			0.60, 0.01	91.9	70.5	87.4	62.1	94.3	75.1	94.0	72.9
	High	.20,.20	0.01, 0.01	74.9	16.6	93.5	16.7	97.6	20.4	93.6	16.8
			0.20, 0.05	83.0	31.1	90.9	30.8	96.0	35.6	92.9	33.0
			0.20, 0.20	83.0	31.2	90.6	30.7	95.9	35.5	92.8	32.8
			0.60, 0.01	91.0	69.5	89.7	65.5	94.6	73.6	93.8	72.4
		.30,.10	0.01, 0.01	73.4	17.1	94.3	17.5	97.4	22.9	93.8	17.7
			0.20, 0.05	82.5	31.4	90.7	30.8	96.0	37.3	93.5	33.4
			0.20, 0.20	82.1	31.4	91.0	30.8	96.2	36.9	93.0	33.1
			0.60, 0.01	90.5	69.7	89.7	64.5	94.0	74.4	93.8	72.8

Figure 1: Boxplot of the distribution of (a) percentage bias and (b) coverage rate for treatment effect estimates on Y_1 , by analysis strategy (CCA, SMI, FMI, MMI) and missingness mechanism (the columns denoted by Ind: individual covariate; Clus: cluster-level covariate, Both and Treat: indicating the variables associated with missingness). The dotted black lines represent (a) no bias and (b) the nominal coverage rate, while the dashed lines represent minimum and maximum acceptable coverage rates.

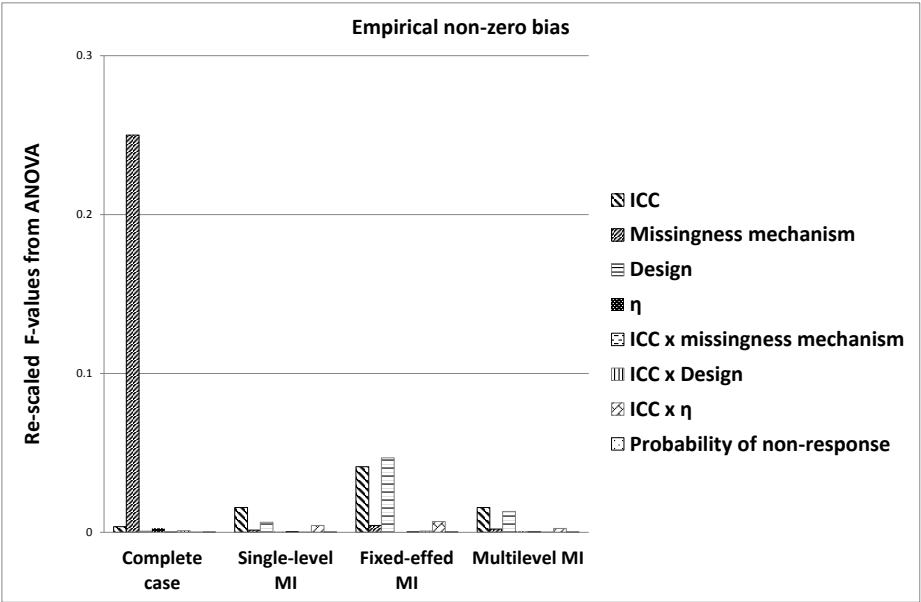


(a) Distribution of Bias

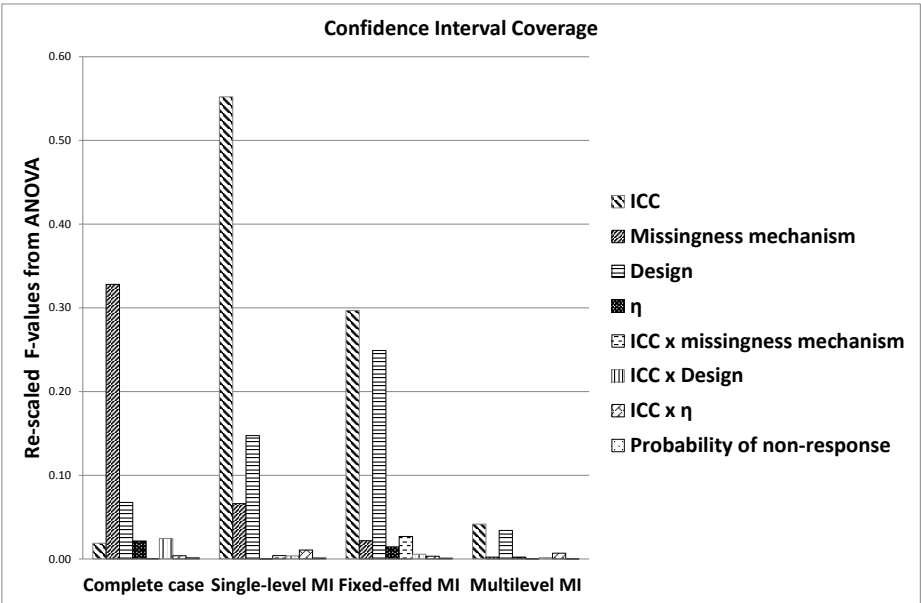


(b) Distribution of Coverage rate

Figure 2: ANOVA results: The height of the bar represents the (normalised) F-statistic value, scaled according to the proportion of scenarios where (a) bias was larger than $1.96\times$ the Monte-Carlo error, and (b) those scenarios where either under or over-coverage are an issue



(a) Bias



(b) Coverage rate

References

- Andridge, R. R. (2011) Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal* **53**, 57-74.
- Campbell, M. K., Fayers, P. M. and Grimshaw, J. M. (2005) Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials*, **2**, 99–107.
- Campbell, M. J., Donner, A. and Klar, N. (2007) Developments in cluster randomized trials and Statistics in Medicine, *Statistics in Medicine*, **26**, 2–19.
- Carpenter, J. R. and Goldstein, H. (2004) Multiple imputation in MLwiN. *Multilevel Modelling Newsletter*, **16**, 9–18.
- Carpenter, J.R., Goldstein, H. and Kenward, M.G. (2011) REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, **45**, 1–14.
- Carpenter, J.R. and Kenward, M.G. (2013) *Multiple Imputation and its Application*. Chichester: Wiley.
- Collins, L.M. and Schafer, J.L. and Kam, C.-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**, 330–351.
- Cornfield, J. (1978) Randomization by group: a formal analysis. *American Journal of Epidemiology*, **108**, 100–102.
- Díaz-Ordaz, K., Kenward, M. G., Coleman, C., Cohen, A. and Eldridge, S. (2014) Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*, DOI:10.1177/1740774514537136
- Donner, A. and Klar, N. (2000) *Design and analysis of cluster randomization trials in health research*. London: Hodder Arnold Publishers.
- Goldstein, H., Carpenter, J.R., Kenward, M.G., and Levin, K. (2009) Multilevel models with multivariate mixed response types. *Statistical Modelling*, **9**, 173–197.
- Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* **60**, 549–576.
- Kenward, M.G. and Carpenter, J.R. (2007) Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, **16**, 199–218.
- Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data. Second Edition*. Hoboken: Wiley.
- Molenberghs, G. and Kenward, M. G. (2007) *Missing Data in Clinical Studies*. Chichester: Wiley.
- Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2011) *MLwiN Version 2.23*. Centre for Multilevel Modelling, University of Bristol.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schafer, J.L (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- Schafer, J.L. (2001) Multiple imputation with PAN. In: *New methods for the analysis of change. Decade of Behavior*. L. M. Collins and A. G. Sayer, Eds, Washington: American Psychological Association. pp. 355–377.

- Schafer, J.,L. and Yucel, R. (2002) Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, **11**, 421–442.
- Taljaard, M., Donner, A. and Klar, N (2008) Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical journal*, **50**, 329–45.
- van Buuren, S. and Groothuis-Oudshoorn, K (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1–67.
- van Buuren, S. (2012) *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.
- White, I. R. and Carlin, J. B (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29** (28):2920–2931.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine* **30**(4), 377–99.
- Yucel, R. and Dermitas, H (2010) Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics and Data Analysis*, **54**, 790–801.